J Mol Evol (1993) 36:182–198

Journal of Molecular Evolution

Statistical Tests of Models of DNA Substitution

Nick Goldman

University Museum of Zoology, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Summary. Penny et al. have written that "The most fundamental criterion for a scientific method is that the data must, in principle, be able to reject the model. Hardly any [phylogenetic] tree-reconstruction methods meet this simple requirement." The ability to reject models is of such great importance because the results of all phylogenetic analyses depend on their underlying models—to have confidence in the inferences, it is necessary to have confidence in the models. In this paper, a test statistic suggested by Cox is employed to test the adequacy of some statistical models of DNA sequence evolution used in the phylogenetic inference method introduced by Felsenstein. Monte Carlo simulations are used to assess significance levels. The resulting statistical tests provide an objective and very general assessment of all the components of a DNA substitution model; more specific versions of the test are devised to test individual components of a model. In all cases, the new analyses have the additional advantage that values of phylogenetic parameters do not have to be assumed in order to perform the tests.

Key words: Phylogenetic inference — Maximum likelihood inference — Evolutionary models — Statistical testing — Hypothesis testing — Molecular clock

All phylogenetic inferences depend on their underlying models. To have confidence in inferences, it is necessary to have confidence in the models. Aligned DNA and RNA sequences are widely used in phylogenetic inference and in this paper I examine models for their evolution by substitution, as used, for example, in maximum likelihood phylogenetic inference (e.g., Felsenstein 1981, 1991b; Bishop and Friday 1985; Hasegawa et al. 1985b, 1987, 1988; Kishino and Hasegawa 1989, 1990). Remarkably few efforts have been made to examine the overall adequacy of common models.

To assess the adequacy of models, they must be incorporated into hypotheses which propose different descriptions of the possible behavior of data and which may be evaluated by objective statistical tests. To perform these tests, a comparative statistic is needed which quantifies the explanatory power of the hypotheses. If the distribution of possible values for the statistic is known, it is possible to distinguish between acceptable and unacceptable deviation from expectations.

If the hypotheses are sufficiently wide-ranging, a good measure of the overall adequacy of models may be obtained. However, it is important to note that a comparison of hypotheses only tests those components in which the hypotheses differ. Any shared assumptions remain untested; if the assumptions are incorrect, this will probably pass unnoticed. Careful choice of hypotheses may eliminate this problem.

The only previous attempts to gain insight into models' overall adequacy appear to be those of Ritland and Clegg (1987), Hasegawa and colleagues (Hasegawa et al. 1988; Kishino and Hasegawa 1990), and Navidi et al. (1991). These approaches make various unverified assumptions: for example, those of Ritland and Clegg and Navidi et al. regarding the distributions of their test statistics.

Present address: Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

At a more detailed level, some work has been performed on testing whether particular summary features of observed data fit the expectations under certain models. For example, the relative proportions of bases (A, C, G, T) in sequences have been tested (Lanave et al. 1984; Gillespie 1986): under most models in common use, the relative proportions should be similar in each sequence studied. Another statistic that has been of interest is the ratio of mean to variance of pairwise counts of differences between sequences (Kimura 1983; Bulmer 1989; Gillespie 1989). These statistics have proved difficult to assess, and it may be inadvisable to concentrate on such specific features of models without first considering the models' overall adequacy.

Further study has been made of methods to distinguish between two (or more) closely related models. In particular, a number of analyses have been proposed to test the equality of rates of substitutional change in different branches of trees (e.g., Langley and Fitch 1974; Wilson et al. 1977; Hasegawa et al. 1990; Kishino and Hasegawa 1990; Hasegawa and Horai 1991). These analyses all assume that the tree relating the sequences studied is known independently, and some assume that the lengths of the branches of the tree are also known. This is generally not the case. Felsenstein (1981) proposed a test of rate constancy which does not make these assumptions, and this is discussed later in this paper. A modified version of Felsenstein's approach is a special case of the new, more general method introduced here. However, by testing only similar models all these methods presuppose that the shared features of the models are suitable descriptions of the processes of DNA evolution. If both (all) models are inadequate, such tests are likely to fail to indicate this.

Hasegawa et al. (1988) and Kishino and Hasegawa (1990) assumed that the numbers of transition and transversion differences in pairwise comparisons of sequences follow a multivariate normal distribution. The means, variances, and covariances of this distribution were assumed equal to the corresponding quantities derived from a continuous-time Markov chain model for DNA substitution (for simplicity, referred to here as a Markov model). The topology of the tree relating the sequences was assumed known and branch lengths were estimated. In such a case, *and providing the assumed tree is derived independently of the subsequent analysis* (not the case if the same data are used to estimate the tree and then subsequently studied by these methods), the probability values reported indicate the goodness of fit of the sequence data to the multivariate normal approximation model (Hasegawa et al. 1988; Kishino and Hasegawa 1990). This method has the disadvantages of requiring an independently known tree and using summarized data. (Pairwise differences only are studied, with a consequent loss of information [Penny 1982].) The adequacy of the multivariate normal approximation is not tested at any stage.

Other work has concentrated on the related but distinct topic of confidence limits (e.g., Felsenstein 1981, 1991b; Hasegawa et al. 1985a,b, 1987; Kishino and Hasegawa 1989). However, the calculation of confidence levels is as dependent on the assumed models as is the inference of phylogeny. Without first the ability to test the adequacy of the models, confidence measures are at best of uncertain value and at worst meaningless. Lockhart et al. (1992) have recently reported a case in which bootstrap sampling has indicated a high level of confidence in a tree that they believe is incorrect. The cause of this disagreement appears to be a failure of the model on which inferences were based. It is vital to tackle the fundamental problem of testing whether particular models are adequate.

In this study, I introduce hypotheses and a test statistic which use likelihood functions based on all the available data to test whether particular phylogenetic models provide a good explanation of observed sequences. Conventional asymptotic approximations to the statistic's distribution are shown not to apply, so significance levels for the test are assessed by Monte Carlo methods (Ripley 1987). The new test is extended to evaluate specific components of models—for example, the assumption of a molecular clock. I demonstrate the use of all these new methods with the analysis of primate $\psi\eta$-globin gene sequences (Koop et al. 1986) and small-subunit RNA sequences (Dams et al. 1988). No investigation is made of the effect of uncertainty in the alignment of sequences, although this itself is recognized as a problem in evolutionary inference (e.g., Thorne et al. 1991, 1992).

Procedures

Hypotheses and Statistics

Suppose we desire to test a particular model of the evolution of DNA, for example, the Jukes and Cantor (1969) model in which all possible substitutions occur at the same rate. A model of current interest will form the null hypothesis (H_0) for the test:

H_0: (a) the sequences are related by an (unknown) phylogenetic "tree" structure
(b) the sites of the sequences have evolved independently, according to the specified model (1)

H_0 is a *composite* hypothesis: the values of parameters of the model, represented by a vector α, say, are not uniquely specified but are to be inferred. For example, α might represent the phylogenetic tree relating sequences, the branch lengths of the tree, and any free parameters of the DNA evolution model. Maximum likelihood methods (Edwards 1972; Felsenstein 1981) can be used to find optimal parameter values ($\hat{\alpha}$) and hence the maximal likelihood value $\hat{L}_0 \equiv L_0(x,\hat{\alpha})$, where x represents the available data, i.e., the aligned DNA sequences. It is convenient to work in terms of the *support* S, the natural logarithm of the likelihood (Edwards 1972), e.g., $\hat{S}_0 = \ln[L_0(x,\hat{\alpha})]$.

The problem remaining is to assess the null hypothesis. An alternative hypothesis that is very general will give the broadest test of the null hypothesis assumptions. The most general hypothesis possible regarding discrete, independent, and identically distributed (i.i.d.) events permits each possible outcome to occur with a fixed probability which is dependent on no parametric model and subject to no restrictions other than the usual laws of probability.

In a problem involving n aligned DNA sequences of length N sites, each site of each sequence will exhibit one of the bases A, C, G, T. Thus the aligned sequences will exhibit at each site one of 4^n possible combinations or *patterns* (Lake 1987; Cavender 1989). If the set of bases is written $\mathcal{S} = \{A, C, G, T\}$, then the set of patterns ($\mathcal{B}$) is $\mathcal{S} \times \mathcal{S} \times \ldots \times \mathcal{S} \equiv \mathcal{S}^n$. The alternative hypothesis (H_1) now takes the form:

$$H_1: \text{P(site } j \text{ exhibits pattern } \mathcal{C} \in \mathcal{B}) = p_{\mathcal{C}}, \forall j = 1, 2, \ldots, N \quad (2)$$

where the probabilities $p_{\mathcal{C}}$ are unconstrained other than that

$$0 \leq p_{\mathcal{C}} \leq 1, \qquad \forall \mathcal{C} \in \mathcal{B}, \qquad \text{and} \quad (3a)$$

$$\sum_{\mathcal{C} \in \mathcal{B}} p_{\mathcal{C}} = 1 \quad (3b)$$

If $N_{\mathcal{C}}$ is defined as the number of sites at which the sequences exhibit pattern $\mathcal{C}$, then under hypothesis (2) the likelihood function is simply

$$L_1 = \prod_{\mathcal{C} \in \mathcal{B}} (p_{\mathcal{C}})^{N_{\mathcal{C}}} \quad (4)$$

This has maximum likelihood solution

$$\hat{p}_{\mathcal{C}} = \frac{N_{\mathcal{C}}}{N} \quad (5)$$

$$\hat{L}_1 = \prod_{\mathcal{C} \in \mathcal{B}} \left(\frac{N_{\mathcal{C}}}{N}\right)^{N_{\mathcal{C}}} \quad (6a)$$

$$\hat{S}_1 = \sum_{\mathcal{C} \in \mathcal{B}} N_{\mathcal{C}} \ln(N_{\mathcal{C}}) - N\ln(N) \quad (6b)$$

As $\hat{S}_1$ depends only on N and the values $N_{\mathcal{C}}$, it is readily calculated for any data set.

Hypothesis (2) will be called the *unconstrained model* or *unconstrained hypothesis*, as no constraints are placed on the probabilities $p_{\mathcal{C}}$ by any phylogenetic model. It is the most general i.i.d. model for the data, and as such permits a very general test for the adequacy of the null hypothesis. Having no components that depend on the evolutionary relationships of the sequences, the unconstrained model allows simultaneous testing of all the evolutionary components of H_0. Rejection of H_0 implies that one or more components are inadequate. It is unlikely that any sequences will fully satisfy the assumption that their sites are i.i.d. but non-i.i.d. models are currently intractable.

The unconstrained model is so general that it includes the possibility of generating the same probabilities that the parametric model of H_0 does. The null hypothesis model is a particular case of the alternative hypothesis (written $H_0 \subset H_1$), and consequently H_1 must give a better explanation of the data in the sense that the likelihood is certain to be higher. However, H_1 is less useful in the sense that it tells us nothing about the evolutionary relationships of the sequences. The difference between the optimal support values for the two hypotheses,

$$\Delta = \hat{S}_1 - \hat{S}_0 \quad (7)$$

can be considered the "cost" of using H_0 to make inferences about phylogeny. A low cost indicates that H_0 is adequate; a high cost indicates that H_0 is performing badly and should be rejected.

Hypothesis Tests

If the maximum likelihood criterion was used to choose between the unconstrained model and any biologically informative null hypothesis, the unconstrained model would always be preferred—even in cases where the null hypothesis was correct. This means that a criterion other than maximum likelihood must be used for comparing models. In traditional statistical theory, a widely accepted statistic for testing the goodness of fit of models is the likelihood ratio statistic $2\Delta = 2\ln(\hat{L}_1/\hat{L}_0) = 2(\hat{S}_1 - \hat{S}_0)$ (Silvey 1975; Lindgren 1976; Kendall and Stuart 1979). When $H_0 \subset H_1$ and the null hypothesis is correct, this statistic is asymptotically distributed

as χ^2_k, irrespective of the true value of any parameters of the null hypothesis (Silvey 1975; Lindgren 1976). The number of degrees of freedom (d.f.), k, is the difference between the numbers of free (estimated) parameters in H_1 and H_0; equivalently, k is the number of restrictions on the parameters of H_1 required to derive the particular case H_0 (Silvey 1975; Lindgren 1976; Kendall and Stuart 1979).

However, a χ^2 distribution cannot be used in the present phylogenetic context for two reasons. Firstly, the χ^2 approximation requires that every attainable category (pattern) be expected to be exhibited a few times (McCullagh and Nelder 1989). This corresponds to well-known rules of thumb for χ^2 tests of Pearson's X^2 statistic for measuring goodness of fit—for instance, that the expected number of observations of each category should be at least five, or that the sample size should be at least four or five times the number of categories:

$$\mathrm{E}(N_{\mathscr{C}}) \geqslant \sim 5, \forall\, \mathscr{C} \in \mathscr{B} \qquad (8a)$$

$$N \geqslant \sim 5|\mathscr{B}| = 5(4^n) \qquad (8b)$$

(see Lindgren 1976). In the problems considered in this paper, such criteria will rarely if ever be satisfied. For most alignments of interest a large proportion of sites will be expected to exhibit one of the four "constant" patterns in which all sequences share the same base. Furthermore, as the number of sequences analyzed, n, increases the number of possible patterns increases very rapidly ($|\mathscr{B}| = 4^n$). Considering these features together, it is apparent that infeasibly long sequences will generally be needed to satisfy conditions similar to equations (8). The effects on the distribution of 2Δ (and similar statistics) caused by the failure to satisfy the rules of thumb are not well known.

To apply a χ^2_k approximation it is necessary to know the appropriate degrees of freedom, k, for the test. While this is straightforward for most conventional tests, it is not at all clear in tree-inference problems. Although the numbers of free parameters in DNA substitution models are simple to count, the assumption that the sequences are related by an *unknown* tree relationship also affects the value of the optimal likelihood, because the estimation procedure selects one tree from the set of all possible trees. The parameterization of tree relationships is complex and not well understood; the vast numbers of possible trees for even moderate-sized problems (Felsenstein 1978), many of them very different from that chosen as optimal, make it clear that the choice amongst trees may have a significant effect on the optimal likelihood. In the asymptotic case that the number of sites N becomes very large, it may be that the tree topology is effectively known and its estimation does not contribute to the d.f. This unusual situation could arise because the set of possible tree topologies is discrete. In practice, however, it is highly unlikely that this limiting case will apply, and it is not yet possible to judge the effect that the choice of trees will have. (See also Felsenstein 1983, 1991a.) Even if an evaluation of this does become available it seems likely that the problem of many patterns not being exhibited would still preclude the use of a χ^2 test.

Ritland and Clegg (1987) and Navidi et al. (1991) did not make allowance for the effect of choosing amongst trees in their tests of the applicability of models. As a result, the estimates of the d.f. for their respective χ^2 approximations are too high. Ritland and Clegg also encountered problems for even small (e.g., five sequence) studies, due to insufficient data. Examples showing the importance of these effects, including the reanalysis of the RNA sequence alignment used by Navidi et al., are given below. These show that the effects are sufficient to render such techniques untrustworthy.

Monte Carlo Tests

When the distribution under H_0 of a statistic S cannot be calculated, it may be possible to use a Monte Carlo statistical test (G.A. Barnard, in the discussion of Bartlett 1963; Ripley 1987). If the null hypothesis model is fully specified (possibly not the case, for nonparametric models), then the probability distribution for data x is known. It is then possible to generate, randomly and repeatedly, data $(x_1, x_2, \ldots, x_m)$ conforming to H_0 and to calculate the corresponding values $(s_1, s_2, \ldots, s_m)$ of the statistic S. A histogram of the values $s_1, s_2, \ldots, s_m$ gives an estimate of the distribution of S. Taking s_1, $s_2, \ldots, s_m$, along with the value s calculated from the original data, gives $m + 1$ values of the statistic S. If the null hypothesis is true, all $m + 1$ values are from the null-hypothesis distribution of S and the probability that s is the rth biggest (or bigger) is simply $r/(m + 1)$. By suitable choice of r and m, a statistical test with a conventional significance level can be constructed. For a 5% test, which I will use throughout this study, it is widely accepted that $m \geqslant \sim 100$ gives good results (Hope 1968; Marriott 1979; Ripley 1987).

A nonparametric estimate of the distribution of S, and a corresponding nonparametric significance test, may be obtained by the related "bootstrap" method (e.g., Efron 1982; Efron and Gong 1983; Efron and Tibshirani 1986). However, it can be of little advantage to base tests of parametric models on less powerful nonparametric methods.

Monte Carlo tests were originally devised for use with *simple* null hypotheses, in which all parame-

ters of the model are assumed known in advance (Ripley 1987). In a phylogenetic context, a simple hypothesis might consist of a known tree relationship between sequences, with all branches of known length, and a fully specified model for DNA substitution. In such a case, there would be no inference being made about parameters of the model, but simply an assessment of a model. In practice, it will be more common to work with a composite hypothesis—the model will generally include an unknown tree and branch lengths, and possibly other parameters to be estimated (e.g., substitution model parameters). This makes straightforward Monte Carlo testing difficult, as it is not clear which parameter values to use for simulation. In this study I use a procedure based on a statistical test devised by Cox (1961, 1962).

Cox's Test

Cox (1961, 1962) considered the question of choice between *distinct* models, i.e., the case in which the hypotheses are not in the nested form $H_0 \subset H_1$ and the χ^2 approximation for the null hypothesis distribution of 2Δ does not apply. If H_0 has likelihood function $L_0(x,\alpha)$ for data x and free (possibly vector-valued) parameter $\alpha \in \Omega_\alpha$, and H_1 has likelihood function $L_1(x,\beta)$ for data x and free parameter $\beta \in \Omega_\beta$, Cox's statistic for discrimination between H_0 and H_1 is

$$\delta = \ln\left[\frac{\sup\limits_{\beta\in\Omega_\beta} L_1(x,\beta)}{\sup\limits_{\alpha\in\Omega_\alpha} L_0(x,\alpha)}\right] = \ln\left[\frac{\hat{L}_1(x,\beta)}{\hat{L}_0(x,\alpha)}\right], \quad \text{i.e.} \tag{9a}$$

$$\delta = \ln[L_1(x,\hat{\beta})] - \ln[L_0(x,\hat{\alpha})] = S_1(x,\hat{\beta}) - S_0(x,\hat{\alpha}) \tag{9b}$$

($-\delta$ is equivalent to l_{fg} of Cox 1961:108, equation 15). In informal terms, this is a likelihood ratio statistic in which the likelihood is maximized separately under each of the two hypotheses, and the ratio of likelihoods formed. Cox's δ is a general form of the statistic Δ discussed above. A large positive value for δ indicates that the data are much better explained by H_1 and that this hypothesis should be accepted in preference to H_0. A large negative value suggests that H_0 is better and should be accepted. The case $H_0 \subset H_1$ is a special case of Cox's test in which $\Omega_\alpha \subset \Omega_\beta$ and necessarily $\delta \geq 0$.

To assess the significance of a particular value of δ it is considered as a random variable Δ, under the null hypothesis. Cox proposed that, since under H_0 the best explanation of the data is that the parameter α takes its maximum likelihood value $\hat{\alpha}$, this value is used for the calculation of the distribution of Δ. This test has been investigated and found satisfactory by Cox (1962) and Lindsay (1974a,b).

Monte Carlo methods can be used for the estimation of the distribution of Δ under $\hat{H}_0$ (Williams 1970; and in the discussion of Atkinson 1970). Data x_i ($i = 1, 2, \ldots, m$) are simulated under $\hat{H}_0$ and for each x_i, δ_i is calculated by maximization of the numerator and denominator of the fractions in equation (9a) [equivalently, of the corresponding terms in equation (9b)] and hence the Monte Carlo distribution of Δ is found. It is crucially important that these likelihoods be maximized independently for each i (i.e., α and β must be estimated separately for each x_i) when performing the Monte Carlo simulation, and not simply calculated under the assumption that $\hat{\alpha}$ and $\hat{\beta}$ [as estimated from the original data; equation (9b)] are correct. In other words, the form

$$\delta_i = \ln\left[\frac{\sup\limits_{\beta\in\Omega_\beta} L_1(x_i,\beta)}{\sup\limits_{\alpha\in\Omega_\alpha} L_0(x_i,\alpha)}\right] \tag{10}$$

must be used, and not $\delta_i = \ln[L_1(x_i,\hat{\beta})/L_0(x_i,\hat{\alpha})]$. This has been stressed by Hall and Wilson (1991), who demonstrate that failure to use the form of equation (10) may have a "profound effect," greatly reducing the power of a test.

Having estimated the distribution of Δ, Monte Carlo testing proceeds as with a simple null hypothesis. For example, if the value of δ obtained from the original data falls in the largest five of the values obtained by including it with 99 simulated values, the hypothesis H_0 is rejected in favour of H_1 at the 5% level. The use of estimated parameters in Monte Carlo simulations has been likened to a "parametric bootstrap" method (Efron and Gong 1983; Efron and Tibshirani 1986). Felsenstein (1988:554) has called such methods "one of the best uses of simulation," contrasting them with simulation studies which strive to infer general properties of analyses from specific examples.

Application to DNA Substitution

In the following DNA substitution model tests, it is in fact the case that $H_0 \subset H_1$ ($\Omega_\alpha \subset \Omega_\beta$). But since a χ^2 test will not be possible and H_0 will be a composite hypothesis, it will not be feasible to test all possible values of $\alpha \in \Omega_\alpha$. The test used will be treated analogously to Cox's test, whose rationale is that if $\hat{H}_0$ is rejected then H_0 is rejected also.

Applying Cox's test to the problem of inferring phylogenies from aligned DNA sequences, $\hat{\alpha}$ repre-

sents the null hypothesis maximum likelihood tree and branch lengths, as well as any free parameters of the substitution process model. Assuming that these values are correct, $\hat{H}_0$ specifies the probabilities with which each possible pattern occurs, i.e., $P(\mathscr{C}|\hat{H}_0)$, for all 4^n patterns $\mathscr{C} \in \mathscr{B}$. Using this multinomial distribution, a sample of N patterns may be drawn randomly. This represents a simulated data set x_1 of n aligned sequences, each of length N sites. By definition, these data conform to the null hypothesis. Calculating $P(\mathscr{C}|\hat{H}_0)$ for all $\mathscr{C}$ may be inconvenient for large n (since $|\mathscr{B}|$ becomes very large). An alternative method for simulating data sets according to $\hat{H}_0$ is to simulate the evolution of a random ancestral sequence, according to the null hypothesis DNA substitution model. This approach is described by Oliver et al. (1989), whose paper also reviews others which have used similar techniques.

The simulated data x_1 are now analyzed under the hypotheses H_0, the appropriate parametric model [hypothesis (1) above], and H_1, the unconstrained model [hypothesis (2) above]. Values of $\hat{S}_0(x_1)$ and $\hat{S}_1(x_1)$ are calculated as though x_1 were the actual data; i.e., likelihoods are maximized independently under the two hypotheses. The difference $\delta_1 = \hat{S}_1(x_1) - \hat{S}_0(x_1)$ gives one value of Δ. If this Monte Carlo simulation is performed repeatedly, many values (δ_i, $i = 1, 2, \ldots, m$) are found; in general,

$$\delta_i = \hat{S}_1(x_i) - \hat{S}_0(x_i) \qquad (11)$$

[from equations (9, 10)]. The values are best summarized in a histogram, and the position of the actual attained value can be compared to this distribution to see whether it is acceptable under the null hypothesis.

The principles of this model testing analysis are not dependent on the particular model used. All that is required is a complete specification of the model so the Monte Carlo simulations can be performed, and the ability to find the maximum likelihood estimates of the parameters of the models. Although particular Markov models are used in this paper (see below), they could readily be exchanged for other Markov models or indeed entirely different DNA substitution models.

Evolutionary Models

Markov Models

In this paper, I study stochastic processes known as continuous-time Markov chains (Cox and Miller 1977) used to model the substitutional evolution of sites of DNA sequences. It may be helpful to summarize some common notation. Each site of a DNA sequence evolving along a branch of a tree is in one of the four states A, C, G, T. Substitutions from a state i to a different state j are assumed to occur according to a Poisson process (Cox and Miller 1977) with instantaneous rate Q_{ij}, and the Q_{ij} form the off-diagonal elements of a square matrix $\mathbf{Q}$ which is completed by setting

$$Q_{ii} = -\sum_{\substack{j \in \mathcal{S} \\ j \neq i}} Q_{ij} \qquad \forall\, i \in \mathcal{S} \equiv \{A, C, G, T\} \qquad (12)$$

This formulation permits the general solution

$$\mathbf{P}(t) = \exp(t\mathbf{Q}) \qquad (13)$$

(Cox and Miller 1977) where the element $P_{ij}(t)$ is the probability that a site is in state j at time $t_0 + t$, given that it was in state i at time t_0, for any $i, j \in \mathcal{S}$ and any $t \geq 0$. If the vector π is the (generally unique) solution to

$$\pi\mathbf{Q} = 0, \qquad \sum_i \pi_i = 1 \qquad (14)$$

then π is the equilibrium distribution for $\mathbf{Q}$ (Cox and Miller 1977). At this equilibrium the mean rate of substitutional change is constant, equalling $-\Sigma\pi_i Q_{ii}$. Rate-time products are confounded in these models, making it impossible to infer anything other than numbers of substitutions unless additional information is available (Felsenstein 1973; Goldman 1990).

Many substitution models are accommodated by this formulation. (See, e.g., Rodríguez et al. 1990.) The Jukes and Cantor (1969) or *equiprobable* ("EQU") model assumes all substitutions occur at the same rate; setting $Q_{ij} = \frac{1}{3}$ for all off-diagonal elements of $\mathbf{Q}$ ensures a mean overall rate of 1. Transition probabilities $\mathbf{P}(t)$ are given by Bishop and Friday (1985:281); the equilibrium distribution is $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

Felsenstein (1981) was dissatisfied with the biological assumption corresponding to the solution $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ of the EQU model, as this suggests that all nucleotides should appear equally often. Felsenstein proposed the use of a model equivalent to $Q_{ij} = \rho_j$ for $i, j \in \mathcal{S}$, $i \neq j$, where ρ_j is the proportion of bases j found in a given set of sequences to be analysed. I refer to this as the "FEL" model; transition probabilities are given by Felsenstein (1981:371) and $\pi = (\rho_A, \rho_C, \rho_G, \rho_T)$. This last formula shows how the equilibrium distribution is matched to the empirically found proportions of the bases.

A further generalization was proposed by Hasegawa et al. (1984). In this case,

$$Q_{ij} = \begin{cases} \kappa\rho_j & \text{if } i \rightarrow j \text{ represents a transition (A} \leftrightarrow \text{G, C} \leftrightarrow \text{T)} \\ \rho_j & \text{if } i \rightarrow j \text{ represents a transversion (A,G} \leftrightarrow \text{C,T)} \end{cases} \quad (15)$$

I refer to this as the "HKY" model after M. Hasegawa, H. Kishino, and T. Yano, the authors who give the full solution $\mathbf{P}(t)$ in Hasegawa et al. (1985b: 163–164). In this case, the equilibrium distribution $\pi = (\rho_A, \rho_C, \rho_G, \rho_T)$, and the equilibrium distribution and transition/transversion rate ratio can be simultaneously controlled by choice of ρ_A, ρ_C, ρ_G, ρ_T, and κ. Values of $\kappa > 1$ make transitions relatively more probable than transversions. The HKY model is very similar to that embodied in Felsenstein's DNAML program (Felsenstein 1991b). The transition/transversion rate ratio R in DNAML can be closely related to κ by the formula

$$R \cong \frac{\kappa(\pi_A\pi_G + \pi_C\pi_T)}{(\pi_A + \pi_G)(\pi_C + \pi_T)} \quad (16)$$

In the implementation of this study, and unlike the DNAML program, the optimal value of the parameter κ is found by maximum likelihood methods. As with the FEL model, the estimates $\hat{\pi}$ are taken from the base frequencies observed in the data. This is done to save computational effort; the estimates are expected to be close to maximum likelihood estimates and an example in which this is so is given below.

Molecular Clock

Kimura, in his "neutral theory" of molecular evolution (e.g., Kimura 1983), has discussed biological models which would lead to rates of DNA substitution being approximately constant over time for the majority of sites of sequences and for different species, i.e., lineages in trees. (See also Bishop and Friday 1985.) This has given a possible explanation of empirical findings of a "molecular clock" (e.g., Pesole et al. 1991). The assumption of a molecular clock corresponds to the assumption that $\mathbf{Q}$ is constant over all branches of a tree. In this paper, I abbreviate models embodying this assumption "SR" to indicate that the Same Rate applies to all branches of a tree. In this case, it is possible to infer the ancestral "root" of a phylogenetic tree (Bishop and Friday 1985; Goldman 1991).

In contrast, Felsenstein (1981) proposed the use of a different overall rate for each branch of a tree. This model does not make the "molecular clock" assumption and corresponds to the assumption that a different scalar multiple of $\mathbf{Q}$ may apply to each branch of a tree. When the molecular clock assumption is relaxed in this way, I use the abbreviation DR to indicate that Different Rates may apply to different branches. In this case, using the Markov models studied in this paper, the root of a tree cannot be uniquely determined (Felsenstein 1981; Goldman 1991). These two possibilities can be used in conjunction with any of the three Markov models EQU, FEL, and HKY described above, yielding models abbreviated SREQU, DREQU, DRFEL, etc. All models assume a tree structure of phylogenetic relationships.

FORTRAN source code of the computer programs used in this study is available on request.

Results

Primate ψη-globin Pseudogenes

As a first illustrative example, I use the ψη-globin gene alignment published by Koop et al. (1986). These sequences represent a pseudogene present in a number of primates. Although longer sequences are now available for more species (Bailey et al. 1991), for simplicity and ease of comparison of results, the sequences for human, chimpanzee, gorilla, orangutan, rhesus monkey, and owl monkey are used here. After alignment and removal of sites where insertions or deletions have occurred, the sequences are 2040 base pairs (bp) long. These data have also been analyzed by Hasegawa et al. (1987, 1988, 1989), Holmes et al. (1989), and Kishino and Hasegawa (1989, 1990).

A test was performed of the adequacy of the SREQU model, i.e., the equiprobable Markov chain model, assuming the same substitution rate throughout the tree (i.e., molecular clock). This gave the following null hypothesis:

H_0: (a) the sequences are related by an (unknown) "tree" structure
(b_1) the sequences have evolved according the "equiprobable" (EQU) Markov model
(b_2) the same overall rate of change applies to all branches of the tree (17)

The tree inferred by maximum likelihood analysis under this model is shown in Fig. 1. The optimal support is

$$\hat{S}_0 = \ln(\hat{L}_0) = -4915.0 \quad (18)$$

To explain these data as well as is possible, while

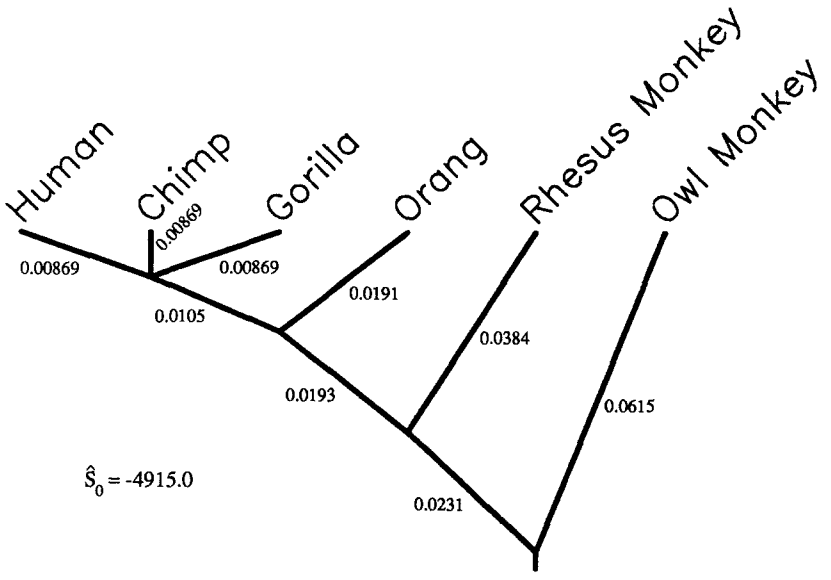

Fig. 1. Maximum likelihood tree for the ψη-globin data under the SREQU model. For SR models, branch-length labels are expressed as relative times when the mean overall rate of substitution is arbitrarily fixed equal to 1. Rooted trees (SR models) are drawn with vertical distances proportional to these times; i.e., the vertical edges of the page form a "time axis." The maximum support $\hat{S}_0$ is as explained in the text.

still assuming the sites of the sequences to be i.i.d., but making no phylogenetic inference, the unconstrained model forms the alternative hypothesis [hypothesis (2)]:

$$H_1: \text{P(site } j \text{ exhibits the pattern of bases } \mathscr{C} \equiv (\mathscr{C}_H, \mathscr{C}_C, \mathscr{C}_G, \mathscr{C}_O, \mathscr{C}_{RM}, \mathscr{C}_{OM}) \text{ for the sequences for human, chimpanzee, gorilla, orangutan, rhesus monkey, owl monkey)} = p_{\mathscr{C}}, \quad \forall j = 1, 2, \ldots, N \tag{19}$$

where the only constraint on the 4096 (4^6) probabilities $p_{\mathscr{C}}$ is that they must form a meaningful set of probabilities; i.e., equations (3) must hold.

From equation (5), the maximum likelihood estimates of $p_{\mathscr{C}}$ are

$$\hat{p}_{\mathscr{C}} = \frac{N_{\mathscr{C}}}{N} = \frac{N_{\mathscr{C}}}{2040} \tag{20}$$

and so under the alternative hypothesis, using equation (6b) the maximum support is calculated to be

$$\hat{S}_1 = \sum_{\mathscr{C} \in \mathscr{R}} N_{\mathscr{C}} \ln(N_{\mathscr{C}}) - 2040 \ln(2040) = -4646.6 \tag{21}$$

Cox's test statistic δ was then calculated [equation (9)]:

$$\delta = \hat{S}_1 - \hat{S}_0 = (-4646.6) - (-4915.0) = 268.4 \tag{22}$$

The significance of δ is assessed by the Monte Carlo test. Monte Carlo data sets were generated by simulating the evolution of random ancestral sequences. For the SREQU model, random ancestral sequences are easily generated according to the equilibrium distribution $\pi = (1/4, 1/4, 1/4, 1/4)$. The optimal null hypothesis tree of Fig. 1 is used with the SREQU model to allow the simulated ancestral sequences to evolve using the appropriate substitution probabilities (Bishop and Friday 1985:281), creating simulated data sets x_i. For each one, $\hat{S}_0(x_i)$, $\hat{S}_1(x_i)$, and thus δ_i [from equation (11)] were calculated.

The histogram of 100 values for δ_i is shown in Fig. 2. It is clear that the actual value of 268.4 for δ falls far beyond the distribution of Δ, and I conclude that the null hypothesis H_0 is false and should be rejected.

Following the rejection of the SREQU model as a description of the evolution of the ψη-globin sequences, it is clearly necessary to alter some or all of the three components of the rejected hypothesis (17). Assumption (a), the existence of a tree structure, is fundamental to phylogenetic studies and will not be questioned in this study. The candidates for alteration are assumptions (b_1) and (b_2), regarding the substitution process model.

In the interests of simplicity, these assumptions are treated separately and the assumption (b_2) that rates of substitution are equal throughout the tree will be retained for the present. An alternative to assumption (b_1) will be considered. Inspection of the aligned sequences (not shown) reveals two relevant features: first, that amongst the constant patterns the data exhibit a bias toward AAAAAA and TTTTTT and away from CCCCCC and GGGGGG. The overall base frequencies also show a prevalence of A and T over C and G. Second, sites at

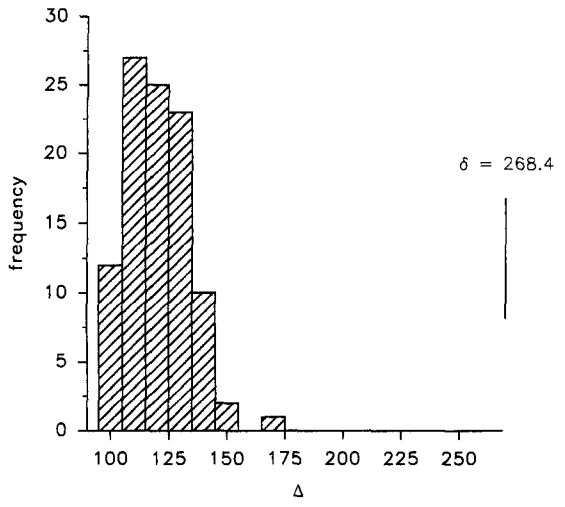

Fig. 2. Monte Carlo distribution of Δ for the Cox test of the SREQU model applied to the ψη-globin data. The attained value of δ falls far beyond the null hypothesis distribution; the SREQU model is rejected.

which one sequence differs from the other five are strongly biased toward this difference being a transition and away from transversions. Considering these points, it would seem that the model could be improved by altering the overall base frequencies so that they better matched the data, and that transitions should be favored over transversions. This suggests the use of the HKY model in which both of these features may be controlled. The null hypothesis for a test of the SRHKY model is:

H_0: (a) the sequences are related by a "tree" structure
(b_1) the sequences have evolved according to the HKY Markov model
(b_2) the same overall rate of change applies to all branches of the tree (23)

The alternative hypothesis, representing the possibility of "any other i.i.d. model," is precisely as before [hypothesis (19)].

Under H_0, the maximum likelihood solution is as shown in Fig. 3. Notice the optimal value of the transition/transversion ratio parameter $\hat{\kappa} = 5.06$ [corresponding to a ratio R = 2.50 in Felsenstein's (1991b) DNAML model: see equation (16) above] indicates a strong bias toward transitions. The optimal support value is:

$$\hat{S}_0 = -4771.5 \tag{24}$$

an increase of 143.5 on the value under the SREQU model. Under H_1, as before, the optimal support is $\hat{S}_1 = -4646.6$. The test statistic δ is now [from equation (9)]:

$$\delta = \hat{S}_1 - \hat{S}_0 = -4646.6 + 4771.5 = 124.9 \tag{25}$$

The significance of this value for δ is tested using a Monte Carlo test. Simulated ancestral sequences of 2040 bp were generated randomly using the empirical base frequencies $(\pi_A, \pi_C, \pi_G, \pi_T) = (0.295, 0.191, 0.237, 0.277)$. The evolution of these sequences along the branches of the null hypothesis optimal tree (Fig. 3) was simulated according to the SRHKY model with the above value for π and $\kappa = 5.06$. The resulting sequences (x_i) were analyzed under H_0 and H_1 and the Monte Carlo sample values of the support difference statistic (δ_i) calculated; 100 simulations give the Monte Carlo distribution $(\delta_1, \delta_2, \ldots, \delta_{100})$ shown in Fig. 4. The attained value $\delta = 124.9$ falls below the 95th percentile of this distribution; i.e., it is not significantly different from the value expected were the null hypothesis true, and the null hypothesis is accepted.

Small Subunit RNAs

As a second example, the new model-testing methods are applied to the small-subunit RNA sequences taken from the alignment of Dams et al. (1988) and used by Navidi et al. (1991) while extending Lake's method of invariants (Lake 1987). The sequences are for *Sulfolobus solfatarius, Halobacterium salinarium* (both Archaebacteria), *Escherichia coli* (Eubacteria), and human (Eukaryota). The aligned sequences contain 1352 bp after sites containing gaps are excluded.

Navidi et al. (1991) analyzed these sequences using maximum likelihood inference based on the DRFEL model and using Lake's method of invariants. Their likelihood analysis differed very slightly from that used in this paper in that the parameter π was estimated by maximum likelihood instead of directly from the observed frequencies of the bases A, C, G, T in the data. For the likelihood analysis, Navidi et al. suggested a likelihood ratio test similar to Cox's test discussed above. They proposed the same alternative unconstrained hypothesis [hypothesis (2)], but neglected to consider the choice amongst all possible trees as a component of the estimation process.

Their simplification suggested the use of a standard likelihood ratio test instead of Cox's test, and Navidi et al. used the χ^2 approximation discussed above. They calculated the degrees of freedom as follows (Navidi et al. 1991:140). The unconstrained hypothesis has 255 d.f. ($4^4 - 1$) and the null hypothesis 8 d.f. [5 independent "branch lengths" (representing expected numbers of substitutions)

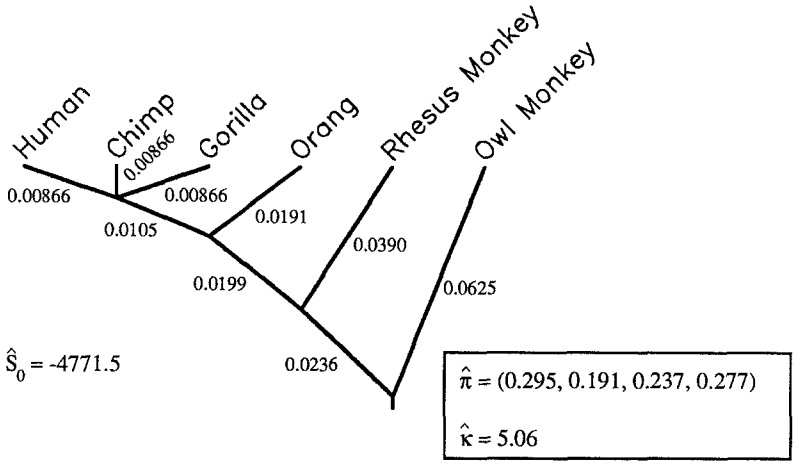

Fig. 3. Maximum likelihood tree for the ψη-globin data under the SRHKY model. Optimal parameter values $\hat{\pi}$ and $\hat{\kappa}$ are as explained in the text.

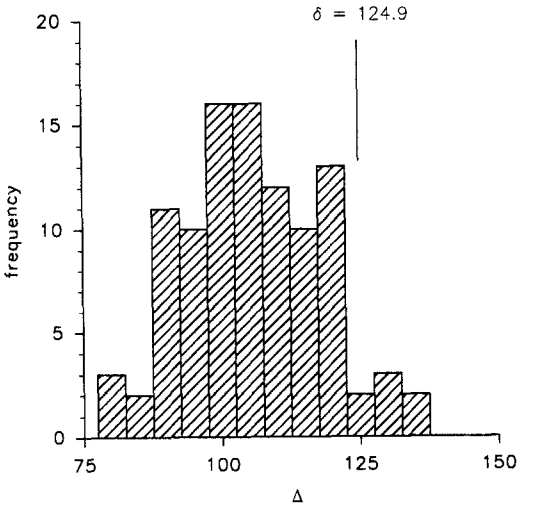

Fig. 4. Monte Carlo distribution of Δ for the Cox test of the SRHKY model applied to the ψη-globin data. The attained value of δ falls below the 95th percentile; the SRHKY model is accepted.

plus 3 parameters π_i (the fourth being fixed by the other 3)]. If the contribution to the null hypothesis d.f. of the estimation of the optimal tree from amongst the set of all possible trees is omitted, the difference (247) should give the d.f. for a χ^2 test, assuming there are sufficient data for the χ^2 approximation to apply. In fact, the DRFEL model has more than 8 d.f. since a choice is also being made between possible trees. Consequently, it would be expected that Navidi et al.'s statistic should have fewer than 247 d.f.

Analysis of these data using the DRFEL model gives the optimal tree shown in Fig. 5a. As expected, this agrees well with the corresponding tree found by Navidi et al. (1991:140, equation 21; Navidi et al.'s branch lengths t_i should all be multiplied by ¾ to match the formulation of this study in which, in contrast to Felsenstein 1981, substitution of a base by itself is precluded). Notice the good correspondence between the empirical estimates of base frequencies, $\hat{\pi}$ of (0.240, 0.254, 0.321, 0.185) in this study and Navidi et al.'s maximum likelihood estimates, r = (0.229, 0.259, 0.309, 0.203). The maximum support value under the DRFEL model is:

$$\hat{S}_0 = -5867.7 \tag{26}$$

the maximum support under the unconstrained model is:

$$\hat{S}_1 = -5591.1 \tag{27}$$

and the Cox test statistic is:

$$\delta = \hat{S}_1 - \hat{S}_0 = -5591.1 + 5867.7 = 276.6 \tag{28}$$

According to Navidi et al.'s simplification, this statistic should be distributed as $\frac{1}{2}\chi^2_{247}$ (i.e., $2\Delta \sim \chi^2_{247}$); its actual distribution, estimated by Monte Carlo methods, is shown alongside a $\frac{1}{2}\chi^2_{247}$ distribution in Fig. 5b. The discrepancy between these distributions is clear. (See below for further discussion.) For these RNA sequences there is no doubt about rejecting the DRFEL model, a conclusion also drawn by Navidi et al. (1991:141).

The data were further analyzed using the SRHKY model, which was found acceptable for the ψη-globin pseudogene alignment discussed above. The optimal tree under this model is shown in Fig. 6a; however, the SRHKY model is also rejected in favor of the alternative unconstrained hypothesis (Fig. 6b).

Further Tests

I now return to the question of choosing between two more-closely related models. By choosing hypotheses that differ only in a small component, a

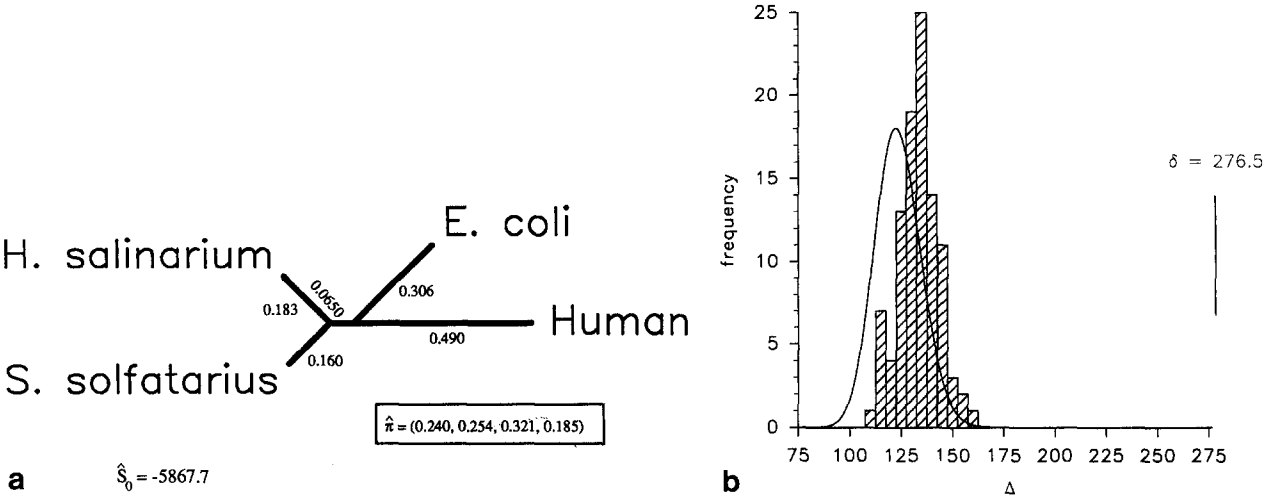

Fig. 5. **a** Maximum likelihood tree for the RNA sequences studied by Navidi et al. (1991) under the DRFEL model. For DR models, branch-length labels are expected numbers of substitutions. Unrooted trees (DR models) are drawn with branch lengths proportional to these numbers. **b** Monte Carlo distribution of Δ for the Cox test of the DRFEL model applied to these data. The attained value δ falls beyond the 95th percentile; the DRFEL model is rejected. The continuous curve is the $\frac{1}{2}\chi^2_{247}$ approximation proposed by Navidi et al. (1991), scaled to be comparable to the histogram, as described in the text.

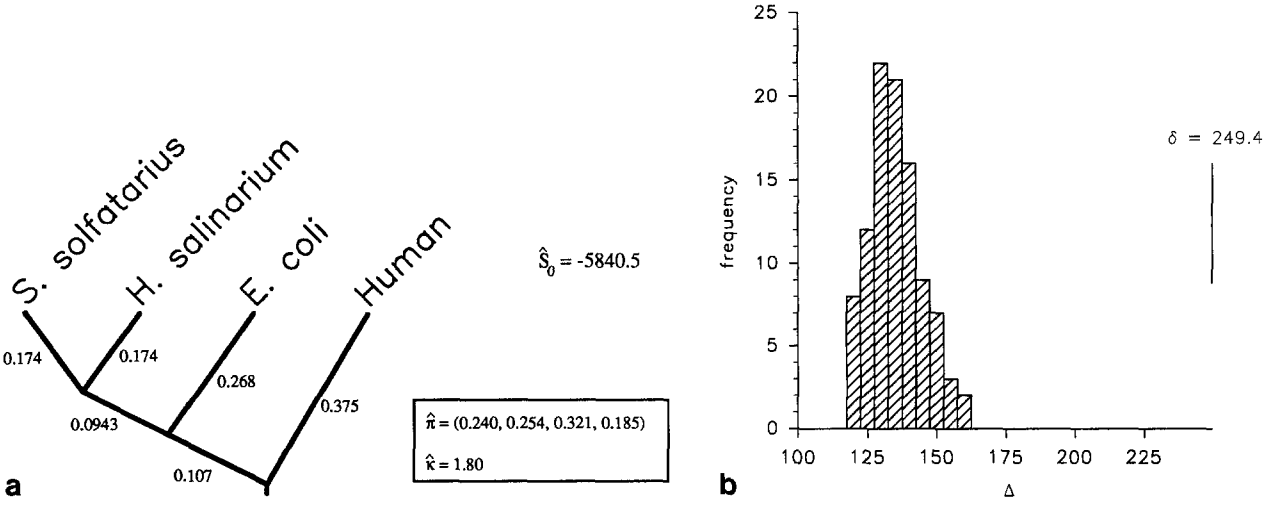

Fig. 6. **a** Maximum likelihood tree for the RNA sequences under the SRHKY model. **b** Monte Carlo distribution of Δ for the Cox test of the SRHKY model applied to these data. The attained value of δ falls beyond the 95th percentile; the SRHKY model is rejected.

test may effectively be performed of the contribution that that component makes to the adequacy of the model. For example, by varying only the molecular clock assumption [e.g., hypotheses ($17b_2$) and ($23b_2$)] it is possible to test whether constancy of substitution rate in all branches of the tree is indicated. Using Cox's test statistic, with significance levels assessed by Monte Carlo simulation, such a test may be performed without making any further assumptions about the relationships of the sequences or other phylogenetic parameters.

The only proposal for a method which avoids untested assumptions about parameter values appears to be that of Felsenstein (1981:374), who suggested the use of a likelihood ratio statistic to compare the assumption of rate constancy throughout a tree (the null hypothesis) against the possibility of different rates in different branches. The significance levels for this test were to be evaluated using a χ^2 distribution and Felsenstein (1981, 1991b) derived the d.f. for such a test, appropriate in the limiting case $N \to \infty$. This test is oversimplified, for some of the same reasons that a χ^2 distribution could not be used to test a single parametric model against the unconstrained alternative hypothesis. In any real problem too many patterns will not be exhibited by the data and, as before, it is unlikely that any asymptotic limit applies. Ritland and Clegg (1987) used Felsenstein's testing methodology and devised other tests between closely related models but faced problems about the sample sizes in relation to the number of possible patterns. Ritland and

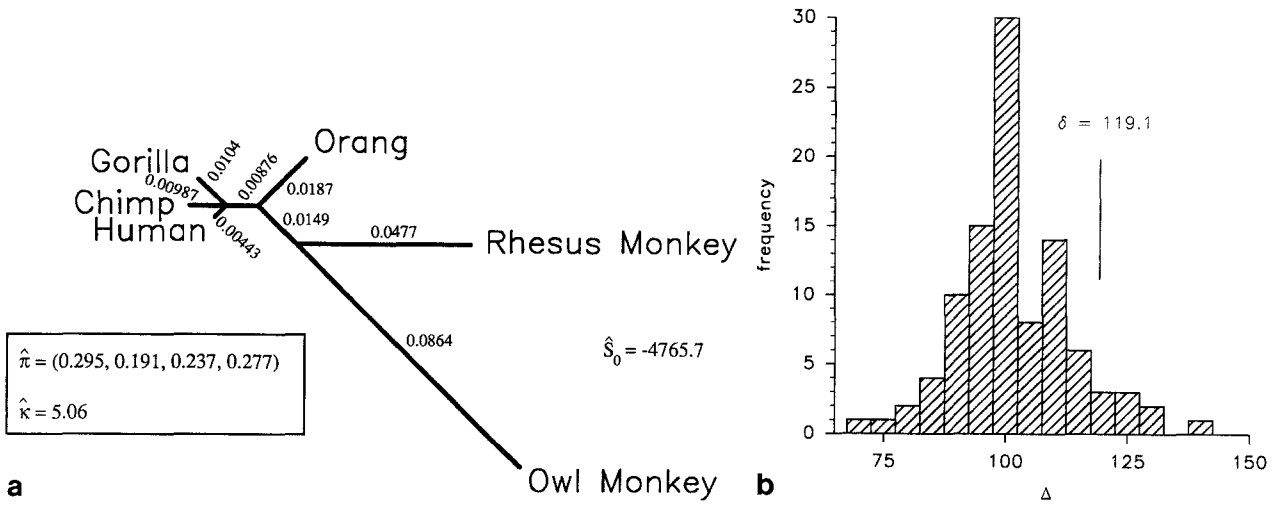

Fig. 7. **a** Maximum likelihood tree for the ψη-globin data under the DRHKY model. **b** Monte Carlo distribution of Δ for the Cox test of the DRHKY model applied to these data. The attained value of δ falls below the 95th percentile; the DRHKY model is accepted.

Clegg (1987:S81) commented on this, but without analysis of the possible effects.

The test I now propose is analogous to the likelihood ratio statistic δ introduced above for testing a parametric model against a very general nonparametric alternative hypothesis. Here, however, interest is in the distinction between two parametric models which may both appear to describe the data adequately. If one model is a particular case of the other, then, as in traditional statistical theory, the null hypothesis will be the simpler (i.e., the less general) hypothesis. This corresponds to the case $H_0 \subset H_1$. Cox's test may still be used when the models are not nested in this way (Cox 1961, 1962). Consequently, the test is more general than the applications shown here might imply and is in fact suitable for comparing any pair of parametric hypotheses. For example, it can be used to choose between two hypotheses which share the molecular clock assumption but embody different Markov models (Goldman unpublished). Its application to the choice between models which respectively do and do not embody the assumption of constancy of overall rate of substitutional change is very similar to the test suggested by Felsenstein (1981), differing only in the method for assessing the significance of the test statistic, although the justification is necessarily somewhat more complicated than Felsenstein's.

Care must be exercised when using this type of test—only the significance of the differences between the hypotheses is evaluated, and any inadequacies shared by the hypotheses will be ignored. Acceptance of a hypothesis by this test means that it is significantly better than the rejected hypothesis, but not that it is necessarily an adequate description of the processes involved.

Consider the primate ψη-globin sequences analyzed previously: the SRHKY model was found to be satisfactory for these sequences. The DRHKY model is a more general version of the SRHKY model, as it permits a different overall rate in each branch of a tree, and thus it necessarily gives a greater optimal support for any given data set. We may desire to know whether this increase is no more than would be expected even if the SRHKY model were actually correct or whether the explanatory power of the DRHKY model is such that it should be accepted. In other words, the test should be between the following hypotheses:

H_0: (a) the sequences are related by a "tree" structure
(b_1) the sequences have evolved according to the HKY Markov model
(b_2) the same overall rate of change applies to all branches of the tree (29)

H_1: (a) the sequences are related by a "tree" structure
(b_1) the sequences have evolved according to the HKY model
(b_2') different overall rates of change apply in different branches of the tree (30)

Since hypotheses (29) and (30) differ only in the assumptions ($29b_2$) and ($30b_2'$) regarding rates in different branches of the trees, only this is tested. Effectively, a test of the molecular clock is performed, without any further assumptions about the models' parameter values having to be made.

The Cox test statistic for this comparison is as in equation (9). Hypothesis (29) has been used previously [hypothesis (23)]; the optimal tree under hypothesis (30) is shown in Fig. 7a. We then calculate:

$$\delta = \hat{S}_1 - \hat{S}_0 = -4765.7 + 4771.5 = 5.8 \quad (31)$$

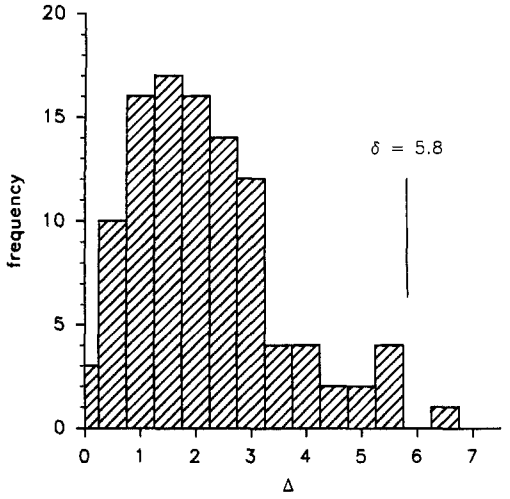

Fig. 8. Monte Carlo distribution of Δ for the Cox test between the SRHKY and DRHKY models applied to the ψη-globin data. The attained value of δ falls beyond the 95th percentile; the null hypothesis (SRHKY) model is rejected in favor of the alternative hypothesis (DRHKY model).

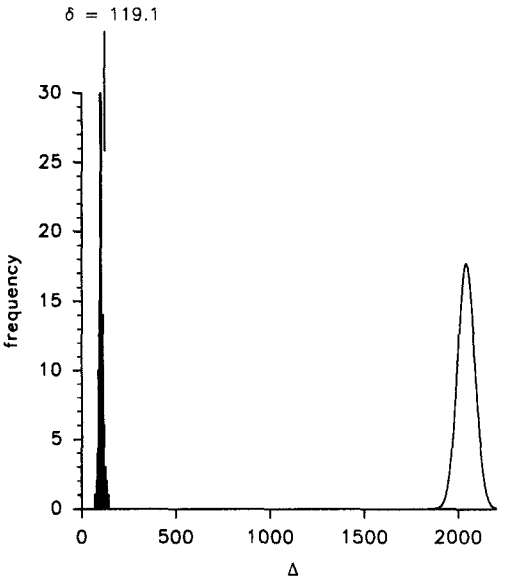

Fig. 9. Monte Carlo distribution of Δ for the Cox test of the DRHKY model applied to the ψη-globin data (left, as Fig. 7a) and the $\frac{1}{2}\chi^2_{4083}$ approximation derived from Navidi et al.'s (1991) method (continuous curve, right).

As usual, the significance of this value is assessed with a Monte Carlo test. Data are simulated under the optimal null hypothesis tree (Fig. 3) and analyzed under hypotheses (29) and (30) to give simulated values δ_i (i = 1, 2, . . ., 100) whose histogram is shown in Fig. 8. This represents the distribution expected for Δ when the null hypothesis is true. As the attained value δ = 5.8 falls beyond the 95th percentile of this distribution, it is judged unlikely to be from the null hypothesis distribution of Δ, and the SRHKY model is rejected in favor of the DRHKY model.

A Monte Carlo test of the DRHKY model against the unconstrained model, using the methods above, confirms that the DRHKY model is accepted as a good explanation of the observed data (Fig. 7b). The conclusion from these results is that two closely-related models both seem adequate parametric descriptions of the evolution of the primate ψη-globin, but one is significantly better than the other. In this case, the test directly comparing the two takes precedence and the DRHKY model should be used by choice. The molecular clock hypothesis is rejected and the preferred estimates of the tree and other parameters are those shown in Fig. 7a.

It is interesting to illustrate the effect of using Navidi et al.'s (1991) χ^2 approximation in the test of the DRHKY model for these data. As noted above, if no account is taken of either the large number of patterns which are rarely (or not at all) exhibited in the data or the selection of one (optimal) tree from amongst the set of all possible trees, then an estimate of the d.f. for a χ^2 approximation can be made. In this example the unconstrained model has 4095 d.f. (4^6 − 1) and the DRHKY model, according to the methods of Navidi et al., has 12 d.f. (3 free parameters π_i + 1 parameter κ + 8 independent branch lengths). Consequently, the estimate of the d.f. for the test is

$$k = 4095 - 12 = 4083 \qquad (32)$$

In the notation of this study, Navidi et al.'s methods estimate the distribution of Δ to be $\frac{1}{2}\chi^2_{4083}$. Figure 9 shows this predicted distribution alongside the simulated distribution of Fig. 7b. A test using the $\frac{1}{2}\chi^2_{4083}$ distribution to assess δ would imply an implausibly good fit of the data to the model; in fact, Fig. 7b shows the fit is much nearer to that expected under the DRHKY model.

As a second example of the test of the molecular clock, the SRHKY and DRHKY models are compared as possible descriptions of the evolution of the small-subunit RNA sequences. The SRHKY model is a particular case of the DRHKY model and so forms the null hypothesis: the null and alternative hypotheses are precisely hypotheses (29) and (30). The SRHKY model has already been used for these data (Fig. 6); the optimal tree under the DRHKY model (not shown) has support $\hat{S}_1$ = −5837.6. Cox's test statistic δ is thus

$$\delta = \hat{S}_1 - \hat{S}_0 = -5837.6 + 5840.5 = 2.9 \qquad (33)$$

The Monte Carlo distribution for Δ for this test is shown in Fig. 10. The attained value of δ is below the 95th percentile, indicating that the apparent in-

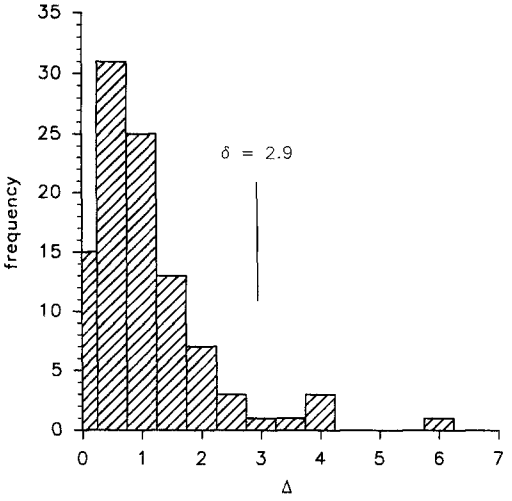

Fig. 10. Monte Carlo distribution of Δ for the Cox test between the SRHKY and DRHKY models applied to the RNA sequences. The attained value of δ falls below the 95th percentile; the null hypothesis (SRHKY) model is accepted in preference to the alternative hypothesis (DRHKY model).

crease in explanatory power of the DRHKY model is in fact no more than would be expected if the SRHKY model was true. Consequently there are no grounds for preferring the DRHKY model: the molecular clock hypothesis is accepted for these sequences. However, this last result must be interpreted with caution. While the SRHKY model is preferred because the more general DRHKY model provides no statistically significant increase in explanatory power, recall (Fig. 6b) that the SRHKY model is itself rejected in favor of the (phylogenetically uninformative) unconstrained model.

Discussion

The phylogenies inferred for the two sets of aligned sequences studied in this paper agree well with the results of previous studies. For the ψη-globin pseudogene sequences, Koop et al. (1986) inferred the same tree relationship as Figs. 1, 3, and 7a, except for the relationships of human, chimpanzee, and gorilla. Koop et al.'s Wagner parsimony analysis suggested the groupings ((human, chimpanzee), gorilla) and ((chimpanzee, gorilla), human) but was unable to distinguish between the two. Holmes et al. (1989) preferred the grouping ((human, chimpanzee), gorilla), applying the methods of Bishop and Friday (1985; equivalent to the SREQU model of this study) and Lanave et al. (1984) to the sequences for human, chimpanzee, gorilla, and orangutan only. Hasegawa and colleagues assumed the grouping (human, chimpanzee, gorilla) to be correct (Hasegawa et al. 1989; Kishino and Hasegawa 1990) and concluded that rates of substitutional change are less in the branches leading to the hominoid sequences. Using the HKY substitution model, they estimated the relative transition/transversion rate parameter $\hat{\kappa}$ ($\hat{\alpha}/\hat{\beta}$, in their notation) to be approximately five. My methods reach broadly the same conclusions as Hasegawa and colleagues, without the assumption of a particular tree topology.

A new result in the present study is that the HKY model is a statistically acceptable description of the evolution of the ψη-globins, with or without the additional assumption of a molecular clock. The development of a new model-testing technique which permits comparison of two closely related models indicates that the molecular clock assumption is not warranted, as the data are significantly better described by different substitution rates. These results are encouraging. The sequences are for a pseudogene which has no coding function, and there is no reason to believe that natural selection operates on them in this respect (Li et al. 1981; Maeda et al. 1988). The sites of such sequences are more likely to evolve independently of one another and according to the same processes, satisfying the i.i.d. assumption.

For the small-subunit RNA sequence alignment, one conclusion drawn is that Navidi et al. (1991) were right to reject the DRFEL model, although their statistical justification for doing so rested upon a potentially dangerous oversimplification. The SRHKY and DRHKY models were also rejected for these data. This is perhaps unsurprising, since for coding sequences the message carried by the DNA is important and any mutations will generally be highly significant. Natural selection will act upon these and will affect substitution. Assumptions of independence and of identical distribution may both be violated in this way. For protein coding sequences, the existence of the genetic triplet coding structure (codons) also imposes further complex biological constraints that are being ignored. Both of these effects have been discussed in detail by Bishop and Friday (1985:296–298). The tree inferred under the SRHKY model (Fig. 6a) and that obtained using the DRFEL model (Fig. 5a) differ from Lake's (1988) assessment of a larger data set using his method of invariants (Lake 1987). Further study of these sequences may enable us to relate these different findings to the failure of the models tested in this study. Most importantly, methods are now available for assessing objectively the suitability of the models upon which inferences are based.

One criticism of phylogenetic analyses has been that most are unable to discriminate between cases in which they are or are not reliable (e.g., Penny and Hendy 1986; Penny et al. 1992). Virtually all

analyses have produced estimates of phylogenetic parameters under certain assumptions but without testing these assumptions, and it has been impossible to assess whether the data are well described by the models or ill described. Without validation of models, it is difficult to assess the worth of the inferences made (Penny et al. 1992). These points may be compared to the stronger statement made by Bross (1990:1214): "Methods that require assumptions that . . . are not verifiable lack statistical validity. Such methods should not be used by biostatisticians to obtain scientific findings."

The computational burden of the simulation tests proposed in this paper is quite considerable. It would be a great advantage to replace Monte Carlo simulation with a suitable approximation for the distribution of Δ. Although, as discussed above, this will be difficult, it may be instructive to try to use as a natural starting point a $\frac{1}{2}\chi^2$ distribution. The expectation of a $\frac{1}{2}\chi^2_k$ distribution is $k/2$ (Lindgren 1976), and using this result it is easy to fit a $\frac{1}{2}\chi^2$ distribution to the Monte Carlo distribution of Δ simply by matching the means. For example, this suggests that the distribution of Δ in Fig. 5b, if it matches any $\frac{1}{2}\chi^2$ distribution, is approximately $\frac{1}{2}\chi^2_{267}$.

Kendall and Stuart (1979) suggest fitting a χ^2 distribution to the actual distribution of a support difference statistic by multiplying the statistic by a constant. In the notation of this study, this corresponds to comparing $c\Delta$ with a $\frac{1}{2}\chi^2_{247}$ distribution in the current example, for some constant c. Future study of the general problem of approximating the distribution of Δ might eventually lead to the reliable use of χ^2 or other tabulated distributions, saving much computational effort.

Notice that in the example of Fig. 5b, the apparent d.f. (approximately 267) is greater than the estimate of 247 made by Navidi et al. (1991), rather than fewer as predicted above. The explanation of this may be that the data fail the criterion that every pattern should be expected to be observed at least a few times. This seems likely to have some effect on the distribution of Δ; in this example it may have increased the apparent d.f. An alternative explanation could be that nonindependence of sites inflates the value of Δ. In contrast, for the $\psi\eta$-globin example of Fig. 9 the apparent d.f. is much less than the naïve estimate of 4083. For this six-sequence problem there are far more candidate trees and we might conclude that the effect on the d.f. of choice between them outweighs any effects of low expected numbers of observations of patterns or low total amounts of data. [Cf. equations (8).]

χ^2 approximations may be more useful in tests between two similar models (McCullagh and Nelder 1989), as discussed above. Further work is needed to evaluate this possibility. In addition, it would be encouraging to show that the distribution of Δ is not strongly dependent on $\hat{H}_0$ (Loh 1985). This is asymptotically the case for traditional problems ($H_0 \subset H_1$) in which χ^2 distributions have been used and would strengthen the argument that if $\hat{H}_0$ is rejected then H_0 is rejected also.

The results of this study remind us how far we are from having widely applicable and accurate models for the evolution of DNA sequences. In cases where specific models are rejected it is of interest to know why—in what respects do the assumptions of the model fail to match the reality of the data? One biological factor missing from the models used in this study is natural selection, which may be manifested as a lack of independence of sequence sites. Attempts to tackle this problem have concentrated on models of inhomogeneity of nucleotide base composition along single sequences. Neighboring base effects have generally been found significant, and some success has been achieved with models in which the base present at a given site depends on those before it (Avery 1987; Bulmer 1987; Churchill 1989). The effects this could have on phylogenetic comparison of many sequences are unknown and could be extremely difficult to ascertain.

Acknowledgments. Adrian Friday gave much encouragement while I performed this work and Tom Kirkwood kindly supplied facilities for the manuscript's completion. Both made helpful comments on various forms of the manuscript, as did Peter Donnelly, Joe Felsenstein, and an anonymous referee. W.C. Navidi made aligned RNA sequences available. This research was performed while I was the holder of S.E.R.C. Studentship Award 88100584.

References

Atkinson AC (1970) A method for discriminating between models. J R Statist Soc B 32:323–345

Avery PJ (1987) The analysis of intron data and their use in the detection of short signals. J Mol Evol 26:335–340

Bailey WJ, Fitch DFA, Tagle DA, Czelusniak J (1991) Molecular evolution of the $\psi\eta$-globin gene locus: gibbon phylogeny and the hominoid slowdown. Mol Biol Evol 8:155–184

Bartlett MS (1963) The spectral analysis of point processes. J R Statist Soc B 25:264–296

Bishop MJ, Friday AE (1985) Evolutionary trees from nucleic acid and protein sequences. Proc R Soc Lond B 226:271–302

Bross ID (1990) How to eradicate fraudulent statistical methods: statisticians must do science. Biometrics 46:1213–1225

Bulmer M (1987) A statistical analysis of nucleotide sequences in introns and exons in human genes. Mol Biol Evol 4:395–405

Bulmer M (1989) Estimating the variability of substitution rates. Genetics 123:615–619

Cavender JA (1989) Mechanized derivation of linear invariants. Mol Biol Evol 6:301–316

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51:79–94

Cox DR (1961) Tests of separate families of hypotheses. Proceedings of the 4th Berkeley Symposium (University of California Press) 1:105–123

Cox DR (1962) Further results on tests of separate families of hypotheses. J R Statist Soc B 24:406–424

Cox DR, Miller HD (1977) The theory of stochastic processes. Chapman and Hall, London, pp 146–198

Dams E, Hendriks L, Van de Peer Y, Neefs JM, Smits G, Vanderbempt I, de Wachter R (1988) Compilation of small subunit RNA subsequences. Nucl Acids Res 16:r87–r174

Edwards AWF (1972) Likelihood. Cambridge University Press, Cambridge, pp 31, 70–102

Efron B (1982) The jackknife, the bootstrap and other resampling plans. Soc Ind Appl Math CBMS—Natl Sci Found Monogr 38

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statistician 37:36–48

Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1:54–77

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

Felsenstein J (1978) The number of evolutionary trees. Syst Zool 27:27–33

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1983) Statistical inference of phylogenies. J R Statist Soc A 146:246–272

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Ann Rev Genet 22:521–565

Felsenstein J (1991a) Counting phylogenetic invariants in some simple cases. J Theor Biol 152:357–376

Felsenstein J (1991b) PHYLIP (Phylogenetic Inference Package) version 3.4, documentation. University of Washington, Seattle

Gillespie JH (1986) Rates of molecular evolution. Ann Rev Ecol Syst 17:637–665

Gillespie JH (1989) Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol 6:636–647

Goldman N (1990) Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst Zool 39:345–361

Goldman N (1991) Statistical estimation of phylogenetic trees. PhD Thesis, University of Cambridge, Cambridge, pp 70–73

Hall P, Wilson SR (1991) Two guidelines for bootstrap hypothesis testing. Biometrics 47:757–762

Hasegawa M, Horai S (1991) Time of the deepest root for polymorphism in human mitochondrial DNA. J Mol Evol 32:37–42

Hasegawa M, Iida Y, Yano T, Takaiwa F, Iwabuchi M (1985a) Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. J Mol Evol 22:32–38

Hasegawa M, Kishino H, Yano T (1985b) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hasegawa M, Kishino H, Yano T (1987) Man's place in Hominoidea as inferred from molecular clocks of DNA. J Mol Evol 26:132–147

Hasegawa M, Kishino H, Yano T (1988) Phylogenetic inference from DNA sequence data. In: Matusita K (ed) Statistical theory and data analysis II. Elsevier, Holland, pp 1–13

Hasegawa M, Kishino H, Yano T (1989) Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. J Hum Evol 18:461–476

Hasegawa M, Kishino H, Hayasaka K, Horai S (1990) Mitochondrial DNA evolution in primates: transition rate has been extremely low in lemur. J Mol Evol 31:113–121

Hasegawa M, Yano T, Kishino H (1984) A new molecular clock of mitochondrial DNA and the evolution of hominoids. Proc Jpn Acad B 60:95–98

Holmes EC, Pesole G, Saccone C (1989) Stochastic models of molecular evolution and the estimation of phylogeny and rates of nucleotide substitution in the hominoid primates. J Hum Evol 18:775–794

Hope ACA (1968) A simplified Monte Carlo significance test procedure. J R Statist Soc B 30:582–598

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism, vol 3. Academic Press, New York, pp 21–132

Kendall M, Stuart A (1979) The advanced theory of statistics, vol 2. 4th ed. Charles Griffin, London, pp 240–252

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, pp 65–89

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Kishino H, Hasegawa M (1990) Converting distance to time: application to human evolution. Meth Enz 183:550–570

Koop BF, Goodman M, Xu P, Chan K, Slightom JL (1986) Primate eta-globin DNA sequences and man's place among the great apes. Nature 319:234–238

Lake JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191

Lake JA (1988) Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331:184–186

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Langley CH, Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. J Mol Evol 3:161–177

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Lindgren BW (1976) Statistical theory. 3rd ed. Macmillan, New York, pp 307–308, 331, 424

Lindsay JK (1974a) Comparison of probability distributions. J R Statist Soc B 36:38–44

Lindsay JK (1974b) Construction and comparison of statistical models. J R Statist Soc B 36:418–425

Lockhart PJ, Penny D, Hendy MD, Howe CJ, Beanland TJ, Larkum AD (1992) Controversy on chloroplast origins. FEBS Lett 301:127–131

Loh W-Y (1985) A new method for testing separate families of hypotheses. J Am Stat Assoc 80:362–368

Maeda N, Wu CI, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1–20

Marriott FHC (1979) Barnard's Monte Carlo tests: how many simulations? Appl Statist 28:75–77

McCullagh P, Nelder JA (1989) Generalized linear models. 2nd ed. Chapman and Hall, London, pp 119, 174

Navidi WC, Churchill GA, von Haeseler A (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 8:128–143

Oliver JL, Marín A, Medina J-R (1989) SDSE: a software package to simulate the evolution of a pair of DNA sequences. CABIOS 5:47–50

Penny D (1982) Towards a basis for classification: the incom-

pleteness of distance measures, incompatibility analysis and phenetic classification. J Theor Biol 96:129–142
Penny D, Hendy MD (1986) Estimating the reliability of evolutionary trees. Mol Biol Evol 3:403–417
Penny D, Hendy MD, Steel MA (1992) Progress with methods for constructing evolutionary trees. TREE 7:73–79
Pesole G, Bozzetti MP, Lanave C, Preparata G, Saccone C (1991) Glutamine synthetase gene evolution: a good molecular clock. Proc Natl Acad Sci USA 88:522–526
Ripley BD (1987) Stochastic simulation. John Wiley and Sons, New York, pp 171–174, 176
Ritland K, Clegg MT (1987) Evolutionary analysis of plant DNA sequences. Am Nat 130:S74–S100
Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142: 485–501
Silvey SD (1975) Statistical inference. Chapman and Hall, London, pp 108–114
Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. J Mol Evol 33:114–124 and Erratum, J Mol Evol (1992) 34:91
Thorne JL, Kishino H, Felsenstein J (1992) Inching toward reality: an improved likelihood model of sequence evolution. J Mol Evol 34:3–16
Williams DA (1970) Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. Biometrics 26:23–32
Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann Rev Biochem 46:573–639

Received April 3, 1992/Revised and Accepted August 13, 1992